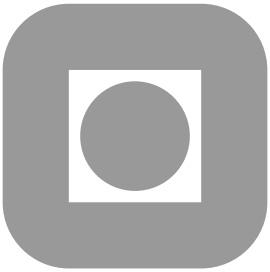
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Prior for flexibility parameters: the Student's t case

by

Thiago G. Martins & Håvard Rue

PREPRINT
STATISTICS NO. 08/2013



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2013/S08-2013.pdf
Thiago G. Martins has homepage: http://www.math.ntnu.no/∼guerrera
E-mail: guerrera@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology,
N-7491 Trondheim, Norway.

Abstract

Extending a basic statistical model can be carried out by using a parametric family of distributions that contains the basic model as a particular case but has an extra parameter, which we denote by flexibility parameter, to control the amount of flexibility or deviation from the basic model. We propose a formal framework to construct prior distributions for the flexibility parameter that place the basic model in a central position within the flexible model, and allow the user to intuitively control the amount of flexibility around the basic model. We propose to assign a prior distribution to the divergence between the basic and the flexible model that encodes the centrality of the basic model within the flexible model. We apply our framework to relax a model based on Gaussian assumptions by extending it to a model based on the Student's t distribution. In this case the flexibility parameter is the degrees of freedom of the Student's t distribution. We show that our priors are robust with respect to its hyperparameters and give sensible results across many different scenarios. We also discuss disadvantages of using priors that do not place the basic model in a central position within the flexible model. Our framework to construct priors for the flexibility parameter is not restricted to the Student's t case and can be applied to a variety of models that can be seen as an extension of a basic model.

Keywords: Priors, Bayesian models, Flexible models

1 Introduction

One elegant way to robustify a model against a particular hazard is to embed the current (basic) model in a more flexible model, indexed by a flexibility parameter. In a Bayesian framework, priors need to be assigned for every parameter in the model, including the flexibility parameter. In this paper we propose a framework to construct prior distributions for the flexibility parameter of a model that have been designed to extend or enrich a specific basic model in order to protect against specific deviations that this basic model might be exposed. Our framework assumes that the basic model plays a central role within the flexible model, and the hyperparameter of our priors controls the degree of flexibility around the basic model. Box and Tiao (1964) applied this same reasoning to propose a prior for a flexibility parameter in the family of power distributions. They regarded their flexibility parameter as a measure of non-Gaussianity and chose a prior distribution with modal value at zero, having adjustable hyperparameters controlling the degree of flexibility of this family of models. We propose a more general approach, where we define the prior distribution on the divergence between the suitably parameterized basic and flexible models for different values of the flexibility parameter in a way that the mode of our prior

distribution for the divergence scale is zero. Then the amount of flexibility will be controlled by the hyperparameter of the prior assigned to the divergence scale.

Some advantages of our approach include 1) *generality*: given the level of abstraction of our framework, the reasoning will be similar whether you want to construct priors for degrees of freedom of the Student's t distribution or precision parameters in the class of intrinsic Gaussian Markov Random Fields, since both can be seen as a flexible model extending a basic one. 2) *robustness* with respect to hyperparameters: priors derived in our framework perform well when data supports the basic model while still being able to capture deviations from the basic model, when indicated by the data. In our examples, we show that commonly used priors for flexibility parameters fail to encode the basic model as the center of the flexible model and that the resulting posterior distributions are much more sensible to misspecifications of its hyperparameters. 3) *Interpretability*: as mentioned, we assign priors to the divergence between the basic and flexible model instead of the original scale of the flexibility parameter. This divergence scale can be used to get a better understanding of the kind of prior knowledge imparted by a given prior, as well as gives a better interpretation for the hyperparameter of our priors.

In this paper, despite the generality of our framework, we have chosen to concentrate on the priors for the degrees of freedom of the Student's t distribution, which is a flexible alternative to a Gaussian distribution. Connections with other model classes will be mentioned in Section 6. Statistical inference based on the widely used Gaussian distribution is known to be vulnerable to outliers, which might imply in a significant degradation of estimation performance (Lange et al., 1989; Pinheiro et al., 2001; Masreliez and Martin, 1977). One possible extension to build a more robust model with respect to outlying observations is to use a mixture of Gaussian distributions (West, 1984, 1987). A particular case of practical relevance is the Student's t distribution, where the degrees of freedom parameter assumes the role of the flexibility parameter and is associated with the degree of robustness of the model. The smaller the degrees of freedom the higher is the robustness of the model against outliers. Besides its property of outliers detection and accommodation, the Student's t distribution has found its way into areas where datasets exhibit heavy tail behavior, as the econometric literature (Chib et al., 2002; Jacquier et al., 1994) for example.

Despite the relevance of the Students't for practical application and the importance of the prior for the degrees of freedom on the statistical analysis, few papers have concentrated on the elucidation of a sensible prior for the degrees of freedom parameter. So far, most of the priors have been proposed on the basis of mathematical tractability for the inference tool at use, as the exponential prior in Geweke (2006), or to facilitate computation of quantities like Bayes

factors, as the uniform prior in Jacquier et al. (2004). More technical work has been done in Fonseca et al. (2008) where Jeffreys priors have been computed for linear regression models with Student's t errors. In Section 4.2 we explain why we chose not to use non-informative priors in the context of model expansion presented here.

As far as we know, a systematic approach that defines weakly informative priors on the divergence scale between two models, where the second model is taken to be an extension around the first have never been presented in the literature. Overall, this idea of using the divergence between two models as a reasonable scale to think about priors dates back to Jeffreys (1946) and have not been properly explored in the literature outside the non-informative or objective prior community. Bayarri and García-Donato (2008) is one good example of work that builds on the early ideas of Jeffreys. They are interested on the derivation of objective priors for computing Bayes factors used to perform Bayesian hypothesis tests. They investigate the ramification of Jeffreys (1961) pioneering proposal and use divergence measures between the competing models to derive the required proper priors, and call those derived priors as divergence-based (DB) priors.

In Section 2 we give a motivation example to illustrate the advantages of using priors for flexibility parameters that were constructed using our framework, while showing the shortcomings of priors that fail to encode the basic model as a reference point for the construction of the flexible model. We present our framework in Section 3 and show the main differences between commonly used priors for the degrees of freedom of the Student's t distribution and our proposed priors in Section 4. We demonstrate the performance of our priors using two examples in Section 5. The first example illustrate our priors in a linear regression model, while the second uses a stochastic volatility model. We present our concluding remarks in Section 6.

2 Motivating example

Consider we want to relate a continuous response $\boldsymbol{y}$ with a continuous covariate $\boldsymbol{x}$. The most simple model that comes to our mind is a standard regression model with Gaussian distributed error terms.

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, ..., n \tag{1}$$

$$e_i \sim N(0, \tau) \tag{2}$$

As mentioned in Section 1, it is well known that the Gaussian distribution is vulnerable to outliers. Therefore, it would be wise to extend this model to a more flexible one by using a

Prior	$\nu = \infty$	$\nu = 5$
exp(0.01)	113.46 (24.60, 426.39)	74.42 (14.07, 319.55)
exp(0.2)	16.99 (8.01, 36.99)	16.36 (9.47, 27.33)
DB(0.3)	83.60 (15.29, 936.05)	19.22 (10.89, 35.77)
DB(0.7)	73.00 (13.45, 808.11)	18.47 (10.02, 33.64)

Table 1: median and $(0.025, 0.975)$-quantiles of the posterior distribution of ν obtained from 50 observation simulated from a linear regression model with Student's t error with ν degrees of freedom. The exponential prior with $\lambda = 0.01$ gives sensible results for the case where $\nu = \infty$, but does poorly for the case where $\nu = 5$. The opposite is true for the exponential prior with $\lambda = 0.2$. Divergence-based (DB) priors give sensible results in both cases, even when we vary the flexibility parameter.

Student's t distribution instead of a Gaussian distribution for the error terms.

$$e_i \sim St(0, \tau, \nu), \quad \nu > 2 \tag{3}$$

Assume we have reparametrized the Student's t distribution in Eq. (3) so that the parameter τ is indeed a precision parameter and is therefore directly comparable with the precision parameter τ in Eq. (2), hence the same notation. We discuss more about parametrization issues in Section 3.1.

We have conducted a simulation study where data is simulated from the regression model of Eq. (1) with the Student's t error terms of Eq. (3). In our simulations we have used $\beta_0 = \beta_1 = \tau = 1$ and have generated 1000 datasets for many different configurations of (ν, n). For each simulated dataset, we have used INLA to perform fully Bayesian inference with different models formed by Eqs. (1) and (3), with the only difference between these models being the prior distribution assigned to the degrees of freedom ν. The priors for all other parameters of the model are all the same across the different models and will be further discussed in Section 5.1, where an extended version of this example will be described.

For now, we focus on Table 1, which illustrates the main motivation of the paper. It shows the median and the $(0.025, 0.975)$-quantiles of the posterior distribution of ν, obtained with four different models under two different scenarios, each with 50 observations. The first column represents the case where data is Gaussian ($\nu = \infty$) and the second column represents a heavy-tailed dataset generated with $\nu = 5$. The first model uses an exponential prior for ν with $\lambda = 0.01$, which has mean 102 (since $\nu > 2$, see Section 3.1). It performs well when the data is Gaussian, attributing high posterior mass for large values of ν. However, it also attributes

high posterior mass for large values of ν in the second scenario, where data is known to have heavy-tails (true $\nu = 5$). The opposite happens when we consider an exponential prior with $\lambda = 0.2$, which has mean 7. It captures well the low degrees of freedom scenario while fail to perform well under the Gaussian data. We conclude that exponential priors for ν are *efficient* in the sense that they perform well under ideal circumstances, but are not *robust* with respect to λ, shown by a significant decrease in performance under less ideal circumstances.

Our framework to build priors for flexibility parameters, to be described in Section 3, aims at constructing priors that are efficient and robust with respect to its hyperparameters. For the Student's t distribution, it means that our priors for ν would imply little loss in performance in case the data is Gaussian while still being able to capture deviations from the Gaussian model in the presence of heavy-tailed data. The last two rows of Table 1 represent the quantiles of the posterior distribution of ν obtained by models where the prior for ν is based on our framework, which we denote by divergence priors, using two different degrees of flexibility, $\text{df} = 0.3, 0.7$. How to construct this so called divergence prior and what is the exact meaning of the degree of flexibility df will become clear in Section 3. The important concept here is that these priors are centered in the basic model, which is a Gaussian model in this context, and df controls how far our model is allowed to deviate from the basic model, so that a divergence prior with $\text{df} = 0.7$ imply in a more flexible model than a divergence prior with $\text{df} = 0.3$, where $0 < \text{df} < 1$. Looking at Table 1, we note that both divergence priors behave reasonably well in both scenarios when compared to the exponential priors, characterizing the robustness with respect to its hyperparameters, as desired.

The purpose of this Section is to emphasize our main objective when designing priors for flexibility parameters, which is efficiency and robustness with respect to the prior hyperparameter. This property is extremely important since there is usually not much prior information available for flexibility parameters, as for the case of the degrees of freedom in a Student's t distribution. A more detailed analysis of the simulation study described in this Section will be given in Section 5.1, where more compelling evidence will be given in favor of the priors developed here.

3 A framework to construct priors for flexibility parameters

3.1 Proper parametrization

We consider a proper parametrization of the model to be the first step in our task of building a sensible Bayesian model. Obviously, the meaning of a proper parametrization depends on the purpose of the model. For example, Cox and Reid (1987) described the situation where a

model $f(y;\boldsymbol{\phi})$ involving an unknown vector parameter $\boldsymbol{\phi}$ of interest is enriched by a nuisance parameter ψ in order to produce a more realistic model. According to them, for this problem to have a clear meaning, $\boldsymbol{\phi}$ should be defined to have an interpretation in some sense independent of ψ. From a Bayesian point of view, besides aid in interpretation, this would allow us to assign similar priors for $\boldsymbol{\phi}$ in the basic model $f(y;\boldsymbol{\phi})$, as well as in the more flexible model $g(y;\boldsymbol{\phi},\psi)$. To be more specific, take $f(y;\boldsymbol{\phi})$ to be a Gaussian density with $\boldsymbol{\phi} = (\mu, \tau)$, where μ is the mean and τ is the precision of the Gaussian distribution and $g(y;\boldsymbol{\phi},\psi)$ to be the Student's t distribution, where the flexibility parameter ψ takes the form of the degrees of freedom ν of the Student's t. Now, if we were to follow the advice of Cox and Reid (1987) we would parametrize the Student's t distribution using $\boldsymbol{\phi} = (\mu, \tau)$ for the mean and precision of the distribution, so that the interpretation of $\boldsymbol{\phi}$ under the Gaussian model would be directly comparable with the interpretation of $\boldsymbol{\phi}$ under the Student's t model.

Although reasonable, this approach is not followed even when the aim of the study is exactly the same as the one exposed here, namely the extension of the Gaussian model through the use of the Student's t distribution to add flexibility and protection against outliers (Lange et al., 1989; Pinheiro et al., 2001). Instead, the more common parametrization of the Student's t is $\boldsymbol{\phi}^* = (\mu, \kappa^2)$ where μ is the mean of the distribution as before and κ is the scale of the distribution. κ is not directly comparable to the precision τ of the Gaussian model, since the precision of the flexible model is given by $(\nu-2)/(\kappa^2\nu)$, and depend not only on the scale parameter κ but also on the degrees of freedom ν. We prefer the former parametrization, specially when we need to assign priors for the parameters, which is facilitated if the parameters have a clear statistical interpretation. For example, we find it easier to think about a prior for the precision of a distribution, which is something we can relate to in a exploratory analysis, than to assign a prior for a scale parameter, which is interpretable only when combined with another parameter in a non-linear fashion.

With that in mind, for the rest of this paper, we are going to parametrize a Gaussian distribution using μ for the mean and τ for the precision, such that $\boldsymbol{\phi} = (\mu, \tau)$ and

$$f(y;\boldsymbol{\phi}) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(y-\mu)^2\right\}.$$

Similarly, the Student's t distribution will also be parameterized by the mean μ, precision τ and flexibility parameter equal to the degrees of freedom, such that $\boldsymbol{\phi} = (\mu, \tau)$ and $\psi = \nu > 2$. We accomplish this by defining

$$y = \mu + \tau^{-1/2} x$$

where $x \sim T_\nu$ is a standardized Student-t with ν degrees of freedom such that its variance is 1

for any value of ν. The p.d.f. of x is given by

$$g(x;\psi) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\left[(\nu-2)\pi\right]^{1/2}}\left(1+\frac{x^2}{(\nu-2)}\right)^{-\frac{\nu+1}{2}}$$

3.2 Prior for flexibility parameters

We advocate a statistical modeling approach where one starts with a basic model, and then start to add more complex structures based on certain diagnostic analysis performed on the results obtained with the initial model. These model extensions are not straightforward, specially within the Bayesian framework where priors need to be assigned for flexibility parameters. In this Section we develop a formal framework to design priors for flexibility parameters in which the desire of having the basic model playing a central role within the more flexible model is properly encoded in the prior distribution. This also illustrates the use of prior distributions to build flexible models in a coherent way, highlighting the advantages of Bayesian statistics for statistical modeling.

We focus in families of parametric models indexed by a flexibility parameter in which the basic model is contained within this family and can be recovered by setting the flexibility parameter to a specific value. This is satisfied in many cases of interest, including the one treated here where a Gaussian distribution is recovered by setting the degrees of freedom $\nu = \infty$ in the Student's t distribution. Our framework is based on two basic assumptions:

1. Since we want the basic model to play a central role within the flexible model, we want our prior distribution of the flexibility parameter to have the mode on the specific value that would recover the basic model.

2. In addition, we want models closer to the basic model to have higher prior density when compared to models that are far away from the basic model.

The two assumptions above are reasonable ones, unless additional prior knowledge states otherwise. For example, looking at Table 1 in Section 2, we see that an exponential prior centered around low degrees of freedom is a reasonable choice if we have a strong prior information that the data comes from a heavy-tailed distribution. However, the results would be disastrous if this prior knowledge turned out to be wrong and the distribution of the data collected were Gaussian or close to a Gaussian. At the same Table 1, we see that the priors designed with the assumptions above behaved well when data were Gaussian (due to assumption 1), and when the data were heavy-tailed (due to assumption 2).

In order to encode the two assumptions of our framework, we propose to assign a prior distribution on the divergence between the standardized basic and flexible models, instead of the more common approach to construct the prior directly on the flexibility parameter scale. For example, let $f(x)$ and $g(x;\nu)$ be the Gaussian and the Student's t densities with zero mean and precision equal to 1. Then we define the divergence d between the Gaussian and the Student's t by

$$d(f(\cdot), g(\cdot;\nu)) = d(\nu), \tag{4}$$

where d is chosen to be a specific divergence measure between the two distributions (see Section 3.2.1). We then define the prior on the divergence d and obtain the prior for the flexibility parameter using the one-to-one relationship $d(\nu)$. We require the prior on the divergence to have mode at zero, which satisfies assumption 1 since $d = 0$ is the value where the flexible model is identical to the basic model. We also require that the prior decays as d increases, so that models closer to the basic model have higher probabilities when compared to models that are far away from the basic model, as stated by assumption 2.

3.2.1 Divergence measures

There is no unique and obvious choice to be used as divergence measure in Eq. (4). However, there are some desirable properties we would like this measure to have. We want d to be non-negative and to vanish if and only if $f(x) = g(x;\nu)$ almost everywhere. This is essential since we want $d = 0$ only for values of the flexibility parameter where the flexible model is identical to the basic model almost everywhere. Finally, d should not depend on the particular parametrization used to describe the densities f and g. This last point is important in order to generate priors which are consistent under one-to-one reparametrization.

Even restricting to divergence measures satisfying the properties above we still have many candidates available. For the results presented in this paper we have used the Kullback-Leibler divergence $\mathrm{KL}(g, f)$ (Kullback and Leibler, 1951), defined by

$$\mathrm{KL}_\nu(g, f) = \int g(x;\nu) \log \frac{g(x;\nu)}{f(x)} dx. \tag{5}$$

Eq. (5) represents the mean information for discrimination between the standardized Student's t distribution $g(x;\nu)$ and the standard Gaussian distribution $f(x)$ provided by an observation from $g(x;\nu)$. The reasoning to use Eq. (5) is that we want to assign more prior mass into values of ν for which it is hard to discriminate between $g(x,\nu)$ and $f(x)$, when data is assumed to come from $g(x,\nu)$. And attribute lower prior mass for values of ν which has greater discriminating power between $g(x,\nu)$ and $f(x)$. The intuition is that if data comes from $g(x,\nu)$ for low values

of ν, where the discrimination power between $f(x)$ and $g(x, \nu)$ is higher (see Figure 1), then the amount of information contained in the data is strong enough to provide a reasonable estimate of ν, despite the lower prior mass. On the other hand, for high values of ν, it is hard to distinguish between $g(x, \nu)$ and $f(x)$ and there is not much information about the tails of the distribution in the data, which makes the model too sensible to prior mass allocated to lower degrees of freedom, leading to poor results in case the prior chosen mistakenly attributes high prior mass for low degrees of freedom.

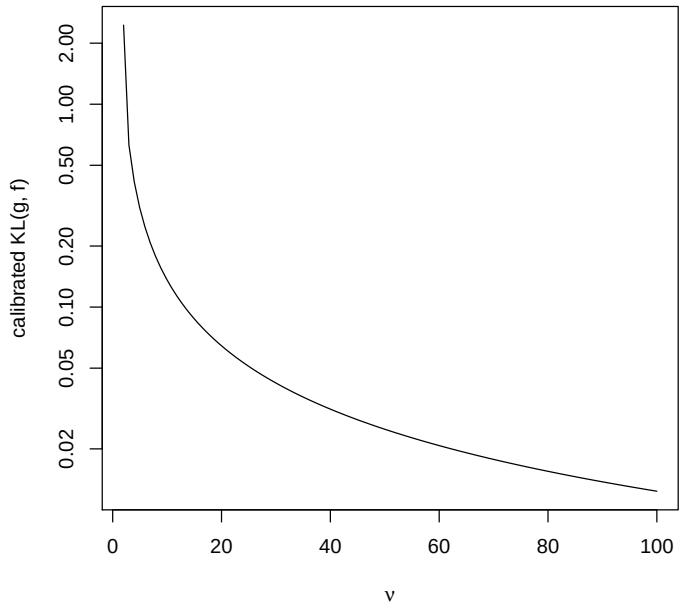


Figure 1: Calibrated Kullback-Leibler divergence between a standardized Student's t with ν degrees of freedom and a standard Gaussian distribution. The y-axis provides equally spaced points in log scale.

Instead of defining the prior directly on Eq. (5), we decided to calibrate the values of $\mathrm{KL}_\nu(g, f)$ according to the Kullback-Leibler divergence between a standard Gaussian and a Gaussian distribution with mean μ and variance 1, which is given by $\mu^2/2$. That is, our prior will be defined on the calibrated divergence given by

$$d(\nu) = (2\mathrm{KL}_\nu(g, f))^{1/2}. \tag{6}$$

There are two main reasons why we use Eq. (6) instead of Eq. (5) as the chosen divergence

metric in our framework. Firstly, it provides a more concrete interpretation and a more intuitive scale for the prior specification. For example, $d(25) \approx 0.05$ means that the Kullback-Leibler divergence between $g(x; \nu = 25)$ and $f(x)$ is approximately the same as the Kullback-Leibler divergence between a standard Gaussian and a Gaussian distribution with mean 0.05 and variance 1. That is, the calibrated divergence d behaves as a location parameter. Secondly, this choice will influence the rate of decay of the induced prior for ν obtained by applying an exponential distribution to the divergence scale, as will be discussed in Section 3.2.2. Figure 1 represents the mapping $d(\nu)$ as defined in Eq. (6) for different values of ν, where the y-axis provides equally spaced points in log scale.

3.2.2 Prior for the divergence

Now that we have decided to use the calibrated KL, defined in Eq. (6), as the divergence measure in Eq. (4), we need to decide on which prior to assign for the divergence d. Remember that, as described in Section 3.2, we want the prior on the divergence $\pi(d)$ to have mode at $d = 0$ (assumption 1), and to decrease as d increases (assumption 2). The exponential distribution

$$\pi(d|\lambda) = \lambda \exp(-\lambda d) \tag{7}$$

satisfies assumptions 1 and 2 and is a reasonable choice. The choice of the exponential prior for d, combined with the use of the calibrated KL divergence leads to a rate of decay of the tail of the marginal prior for ν when $\nu \to \infty$ that are similar to the rate of decay of the independent Jeffreys prior developed by Fonseca et al. (2008), which is $O(\nu^{-2})$. This tail behavior can be seen in Figure 2, which plots the tail behavior for our divergence-based priors with $df = 0.3, 0.5, 0.7$ and 0.9 (dashed lines) together with the tail behavior of the independent Jeffreys prior (solid line) for ν. From Figure 2, we can also see how close the divergence-based prior with df $= 0.7$ is to the independent Jeffreys prior for ν. This similarity will also be illustrated in Figure 6 when we discuss non-informative priors in Section 4.2.

A nice feature of this prior specification is that the λ parameter have an interpretation that provides an easy and intuitive mechanism for the user to increase or decrease the flexibility of the prior. For example, if we set

$$\lambda = -\log(p)/d(\nu^*) \tag{8}$$

it means that the prior attributes $100p\%$ of its probability mass on the interval $(2, \nu^*]$, where $d(\nu)$ is defined in Eq. (6). For example, if we set $\lambda = -\log(0.3)/d(10) \approx 8.85$ our prior will have 30% of its probability mass on the ν interval $(2, 10]$. However, if we have an indication that there is a higher chance of heavy tails we might want to increase the prior probability on low values of

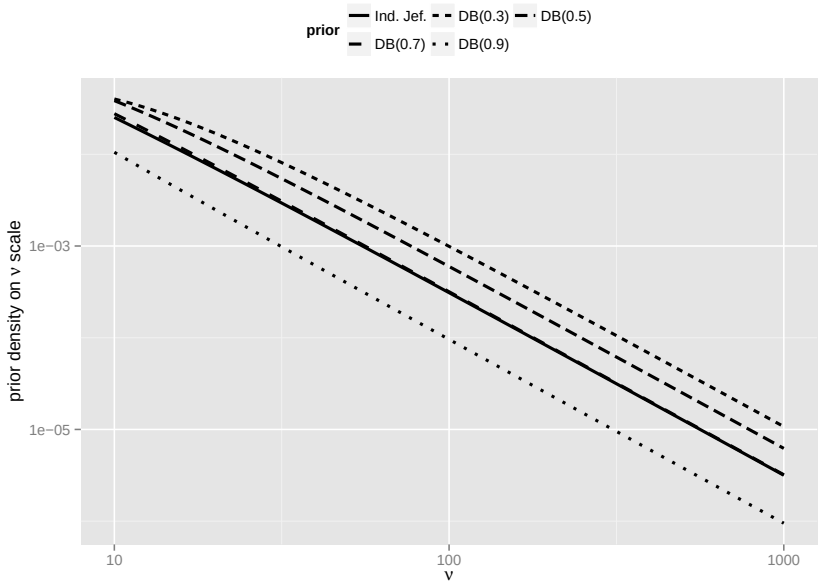
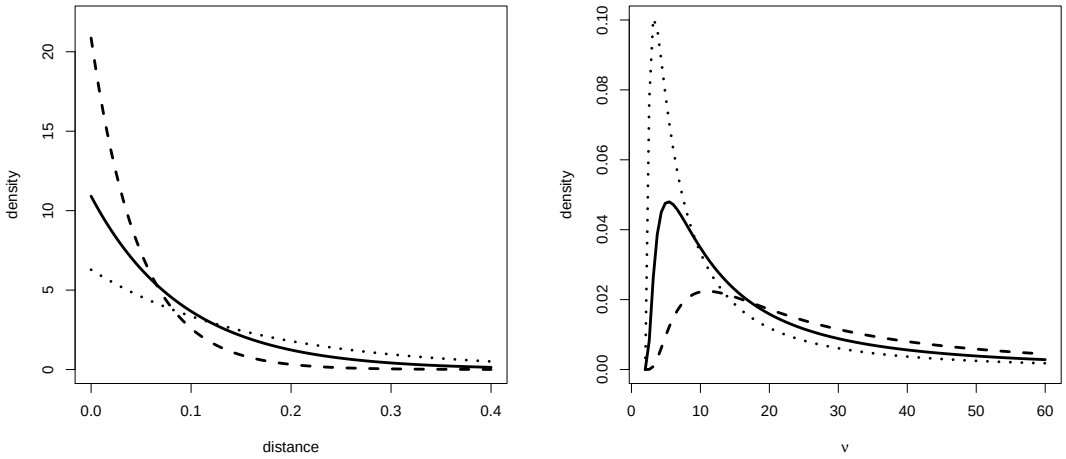


Figure 2: Tail behavior of different marginal prior distributions for ν: From top to botton we have divergence-based (DB) prior with df = 0.3, DB with df = 0.5, DB with df = 0.7, independent Jeffreys prior and DB with df = 0.9. Both the x- and y-axis are plotted in log scale.

ν and can for example decide to assign 50% prior probability on the ν interval $(2, 10]$ by setting $\lambda = -\log(0.5)/d(10) \approx 5.10$. The same is true if we expect just minor deviations from the Gaussian assumptions and decide to set $\lambda = -\log(0.1)/d(10) \approx 16.93$ to have 10% prior mass for ν in $(2, 10]$. Notice that although this prior specification allows the kind of flexibility just described, it does that while still satisfying both assumptions of our framework. We can see that in Figure 3(a), where we show our prior density on the divergence scale for different values of λ. We can see that even though we change λ we have the mode of the prior on $d = 0$ (assumption 1) and a decreasing density as d increases (assumption 2) for all three cases displayed. We will see that this is not true for other priors for the degrees of freedom used in the literature and can be a source of problems on those cases. Figure 3(b) shows the implied priors on ν scale after applying the Jacobian of the mapping $d(\nu)$. Later when comparing with other priors used in the literature, we note that the priors on ν scale obtained by our framework have heavier right tail, which is expected since our priors have encoded the special role of the Gaussian assumption in our flexible model.



(a) (b)

Figure 3: Prior specification using our proposed framework with an exponential prior on the divergence scale (a) and the implied prior on the degrees of freedom scale (b). We have $P(\nu < 10) = p$, for p = 0.1 (dashed), p = 0.3 (solid) and p = 0.5 (dotted)

4 Commonly used priors and related work

4.1 Commonly used priors on the divergence scale

Our approach proposes to specify priors for flexibility parameters on the divergence between the basic and flexible models. Besides being useful to encode assumptions 1 and 2 of our framework, the divergence scale is also useful to try to understand what commonly used prior distributions for flexibility parameters, usually elicited directly on the flexibility parameter scale, means in terms of the relation between the basic and flexible model. From this point of view, some of the priors that seem reasonable at first, might actually be considered inappropriate for specific cases.

For instance, the uniform prior for the degrees of freedom was advocated in Jacquier et al. (2004) without any specific justification for this choice besides that it facilitated the computation of Bayes factors when using the estimation algorithm proposed by the authors of that paper. Another case where the uniform prior for the degrees of freedom gets highlighted is on the examples page[1] of the OpenBUGS software (Lunn et al., 2009), where an uniform prior for ν is used to learn about the degrees of freedom from simulated data. Figure 4 shows the implied

[1] http://www.openbugs.info/Examples/t-df.html

priors on the divergence scale when uniform priors are used for ν. The solid line in Figure 4 refers to the prior on the divergence scale obtained from a $U(2, 40)$ on ν, similar to what was used in Jacquier et al. (2004), while the dashed line is obtained from a $U(2, 100)$, which was used as a prior for ν on the OpenBUGS example mentioned earlier.

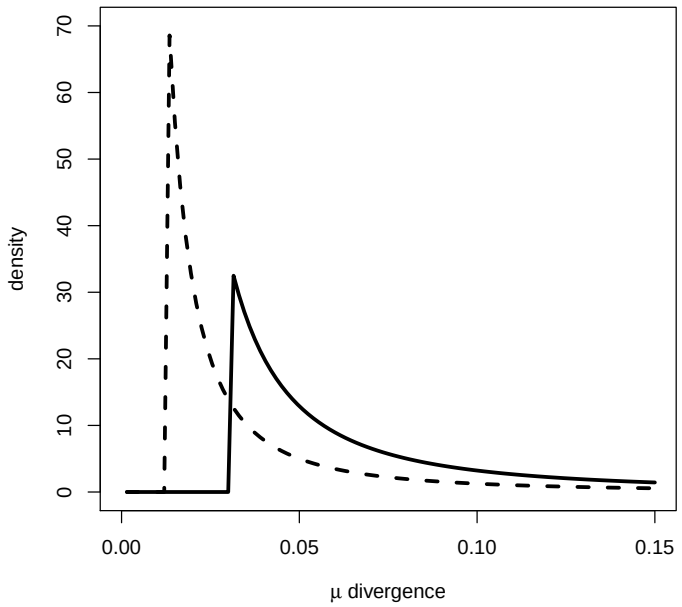


Figure 4: Implied prior distribution on the divergence when an uniform prior distribution is assigned for ν. Solid line is obtained from a $U(2, 40)$ while dashed line is obtained from a $U(2, 100)$.

There are two interesting properties of the uniform distribution that gets highlighted when looking at Figure 4. The first one is that no prior mass is allocated for the Gaussian model (i.e. for $d = 0$), which seems obvious after you realize it, since the Gaussian model is recovered only when $\nu = \infty$. However, this feature happens for any finite upper bound on the uniform distribution, which is not a desirable property when our intention is to extend a basic model. The least you could expect is to be able to recover the basic model if the data says so, which is denied when using an uniform distribution for ν as illustrated by the absence of prior probability mass around $d = 0$ in Figure 4. The second property, which is less obvious when looking the prior on the degrees of freedom scale, is that most of the prior probability mass is concentrated in models that are close to the Gaussian model, and it gets more concentrated as we increase

the upper bound of the uniform distribution. This behavior is very well illustrated in Figure 4 where we can see that the $U(2, 100)$ (dashed line) is more concentrated in the region close to $d = 0$ than the $U(2, 40)$ (solid line). To be more specific, the $U(2, 100)$ attributes only 28.6% of its probability mass for ν on the interval $(2, 30)$, and this number goes to 14.1% if a $U(2, 200)$ is used as a prior for ν, a fact that sometimes goes unnoticed if we only plot the prior on the degrees of freedom scale. To sum up, when we use the uniform prior for ν we get a contradictory behavior where we are unable to recover the basic model since we assign zero probability mass at $d = 0$, while at the same time we concentrate most of the prior probability mass on models that are close to the Gaussian, and this concentration gets more and more dense as we increase the upper bound parameter of the uniform distribution.

An exponential prior for ν was suggested by Geweke (2006), mainly due to its mathematical tractability within a Gibbs sampling framework. Although it might look a reasonable first choice, we see an interesting pattern when we look at Figure 5, which shows the implied prior distributions on the divergence scale obtained from different exponential priors assigned directly on the ν scale. If we start with the prior implied by an exponential with rate parameter $\lambda = 0.01$, represented in Figure 5 by the solid line, we see that this prior concentrates most of its prior mass on models close to the Gaussian model. Specifically, it attributes only 24.4% of its probability mass for ν on the interval $(2, 30)$ and only 7.7% on the interval $(2, 10)$. Besides, although it has non-zero prior density on $d = 0$ its mode is slightly positive. If we then needed to add more flexibility on the model, i.e. assign more prior mass for low values of ν we would have to choose a higher rate parameter. Figure 5 also shows the implied priors on the divergence obtained from a exponential distribution for ν with rate parameters equal to 0.1 (dashed), 0.5 (dotted) and 1 (dot-dashed). Note that as we increase the rate parameter of the exponential distribution more prior mass is assigned to low values of ν, i.e. to high values of the divergence d. However, as we attribute more prior mass to models with low degree of freedom, we do this by neglecting the models that are closer to the Gaussian model, which is represented in Figure 5 by the change in the mode and by the absence of probability mass for $d = 0$ and on its vicinity. This is an undesirable behavior if we expect to construct a model that extends the model based on the Gaussian assumption, without excessively favoring non-Gaussianity a priori. This behavior can be contrasted with the one exhibited by our proposed priors on the divergence as illustrated in Figure 3(a). There we see that we are able to increase the prior mass for models that are further away from the Gaussian without changing the decaying effect of the prior, which is desirable for our purposes, as described in Section 3.2. As we have seen in Section 2, the exponential distribution will do well if we have prior knowledge about the adequate degrees of freedom for

the analysis, in which case we can use the exponential distribution to concentrate the prior mass around this specific value. However, misleading results will be obtained if the appropriate degrees of freedom actually differs from what was believed to be reasonable a priori.

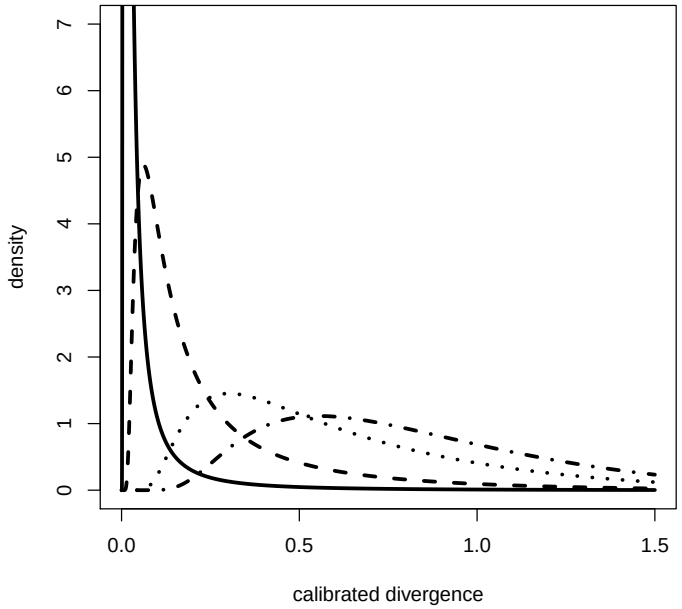


Figure 5: Implied prior distribution on the divergence when an exponential prior distribution with rate parameter equal to 0.01 (solid), 0.1 (dashed), 0.5 (dotted) and 1 (dot-dashed) is assigned for ν.

4.2 Non-informative priors

Another road to take when thinking about priors for flexibility parameters is to use non-informative priors. One example is the Jeffreys prior for ν that was developed by Fonseca et al. (2008) for the specific case of a linear regression model with Student's t error. We do not use non-informative priors in the context of model expansion described here for two reasons. Firstly, the main idea behind non-informative priors is the (hope of) absence of prior knowledge. This is not exactly what we want, as stated in the two assumptions of Section 3.2. Our proposed approach includes some generic information on the prior that are reasonable for a wide range of applications in which the intention is to build more flexible models that have the basic model as its basis. We think this is along the lines of weakly informative priors defended

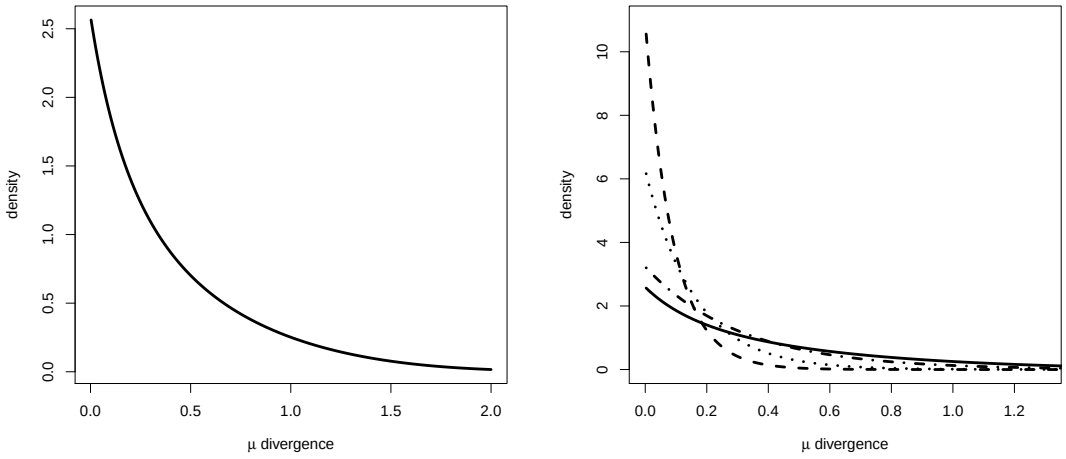
by (Gelman, 2006; Gelman et al., 2008) where we use a minimal prior knowledge to set proper priors for the flexibility parameter that produces stable, regularized estimates, while still offering enough flexibility for inclusion of extra knowledge. Secondly, non-informative priors depend on the model structure and should be recomputed if any change happens to the structure of the model. This is unfortunate because we often need to change our model structure when analyzing data, as for example when we decide to add and/or modify temporal/spatial effects within a generalized additive model (GAM) framework. Besides, the computations required to compute the non-informative priors gets very complicated as the complexity of the model grows, since those computations usually depend on information matrix that are sometimes not even available in closed form.

Notice that we have given two reasons why we do not use non-informative priors in the context of this paper. We have by no means said that non-informative priors are unreasonable or that it should not be useful in other frameworks. Figure 6(a) shows how the independent Jeffreys prior (Fonseca et al., 2008)

$$\pi(\nu) \propto \left(\frac{\nu}{\nu+3}\right)^{1/2} \left\{\psi'\left(\frac{\nu}{2}\right) - \left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right\}^{1/2}$$

behave on the divergence scale. The independent Jeffreys prior above was computed by Fonseca et al. (2008) in the context of a linear regression model with Student's t error by assuming that the marginal prior of the fixed-effects $\boldsymbol{\beta}$ and the joint prior of the scale and degrees of freedom (σ, ν) are independent a priori. Then the prior for each of this groups of parameters are computed by assuming the parameters outside the group as fixed and applying a Jeffreys-rule prior, which is given by $\pi(\theta) \propto \sqrt{\det I(\boldsymbol{\theta})}$, where $I(\boldsymbol{\theta})$ is the Fisher information matrix and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \nu)$. Notice that the grouping of parameters to which independence is assumed is subjective, and changes to this assumption imply in different prior distributions.

Looking at Figure 6 and comparing with the priors analysed in Section 4.1, the independent Jeffreys prior as defined by Fonseca et al. (2008) is in agreement to what we propose in our paper, and to what we call reasonable within our context, i.e. more prior mass close to the basic model and a decaying effect as the divergence increases. Figure 6(b) plot the independent Jeffreys prior (solid) together with exponential distributions that assign 30% (dashed), 50% (dotted) and 70% (dot-dashed) prior mass for ν in the interval $(2, 10)$. Notice how close the independent Jeffreys prior is to the exponential that assign 70% of prior mass for ν in the interval $(2, 10)$. One way to interpret this is that the independent Jeffreys prior for ν is similar to our proposed priors with a high degree of flexibility df, which might not the the most appropriate choice as illustrated by Figure 9 in Example 5.1.



(a) (b)

Figure 6: (a) Implied prior distribution on the divergence when an independent Jeffreys prior distribution is assigned to ν. (b) Implied prior distribution on the divergence when an independent Jeffreys prior distribution is assigned to ν (solid) together with our proposed exponential distribution on the divergence scale that satisfy $P(\nu < 10) = p$, for p = 0.3 (dashed), p = 0.5 (dotted) and p = 0.7 (dot-dashed).

5 Examples

5.1 Robust linear regression models

Here, we continue with the simulation example of Section 2, where we compare robust linear regression models. The only different between the models used in the comparison is the prior distribution assigned to the degrees of freedom ν of the Student's t distribution of the error terms. The models are formed by Eqs. (1) and (3). Independent Gaussian priors with large variance are assigned to the fixed-effects β_0 and β_1. In order to focus on the estimation of the degrees of freedom parameter, we have used an informative gamma prior for the precision τ, with shape and rate parameters equal to 1 and 1.06 respectively, which were chosen following guidelines described in Fong et al. (2010).

As mentioned in Section 2, we have generated 1000 datasets for many different configurations of (ν, n) with $\beta_0 = \beta_1 = \tau = 1$, where n is the sample size of each dataset. Figure 7 illustrates results obtained from six different models for $n = 100, 1000$ and $\nu = 5, 20, \infty$. For each scenario, the first three intervals (solid line) in Figure 7 are generated by the models with divergence priors

for ν with df $= 0.3, 0.5, 0.7$, respectively. The last three intervals (dotted line) are obtained from models with exponential priors for ν with rate parameter $\lambda = 0.01, 0.05, 0.2$, respectively. Each interval is the result of taking the median of 1000 $(0.025, 0.5, 0.975)$-quantiles estimates obtained from a specific model fitted to the 1000 datasets of a specific scenario. For example, the first interval in the top left panel of Figure 7 represents the median of the 1000 $(0.025, 0.5, 0.975)$-quantiles estimates obtained from fitting the robust linear regression model with the divergence prior for ν with df $= 0.3$ to the 1000 datasets generated with $n = 100$ and $\nu = \infty$.

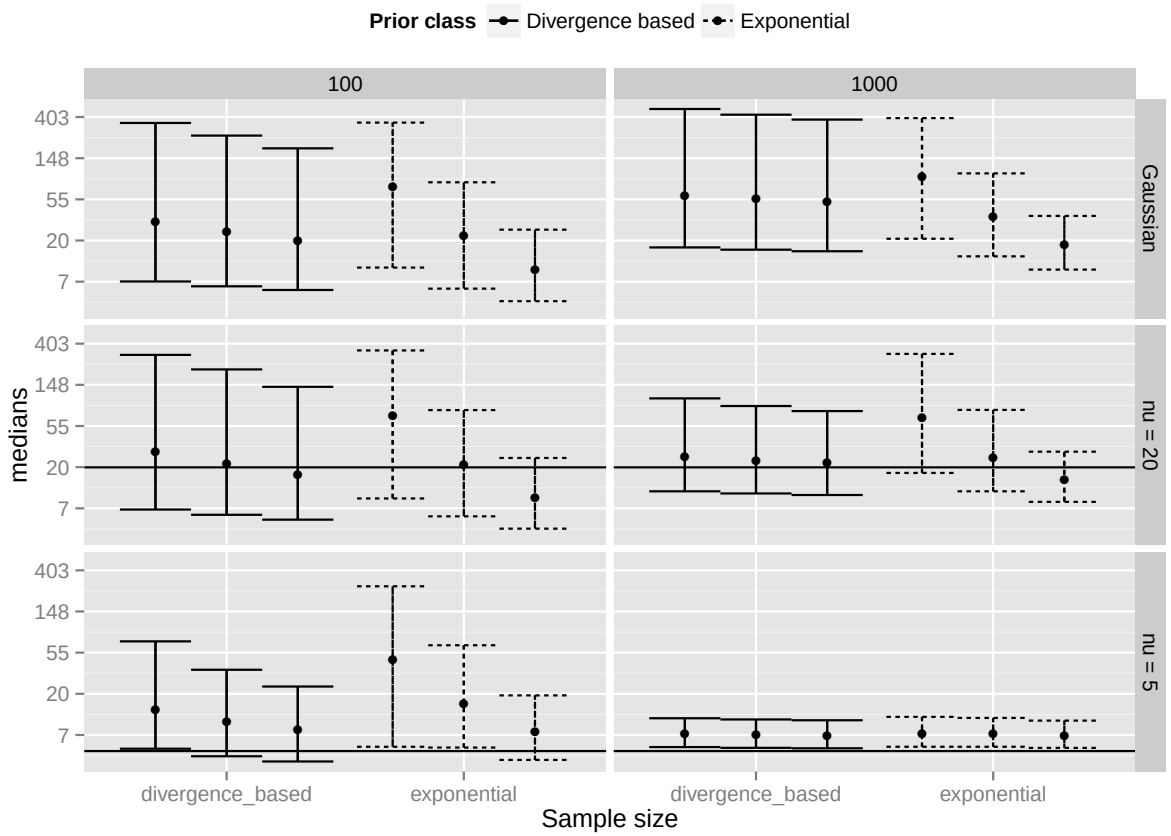


Figure 7: Median of the 1000 $(0.025, 0.5, 0.975)$-quantiles estimates obtained from fitting the robust linear regression model with different priors for ν for sample sizes $n = 100, 1000$ and degrees of freedom $\nu = \infty, 20, 5$. The first three intervals in each scenario (solid line) correspond to our divergence-based priors with df $= 0.3, 0.5, 0.7$ respectively. The last three intervals in each scenario (dotted line) correspond to exponential priors with rate parameter $\lambda = 0.01, 0.05, 0.2$

Figure 7 allow us to arrive at similar conclusions obtained in Section 2, but with more

details. We can think of Figure 7 as a 3×4 matrix. The first row represents the Gaussian scenario ($\nu = \infty$). If we look at the divergence priors with moderate number of data points ($n = 100$), we see that all three priors in this category provide intervals that includes high values of ν, which is good, given the Gaussian data. Moreover, when we increase the number of data points to $n = 1000$, these three models have now more information and the use of this information is represented by the wider intervals including even higher values of ν. Now, if we look at the models with exponential prior for ν the results are surprising. In the presence of Gaussian data, the posterior inference with respect to ν will be highly dependent on the rate parameter λ chosen for the exponential prior. See for example that the results for $n = 100$ changes abruptly if we choose $\lambda = 0.01$ or $\lambda = 0.2$. In addition, this situation remains the same even with the increase of data points to $n = 1000$, as if the prior was completely dominating the data. As a matter of fact, notice that the intervals obtained with the exponential distribution for ν looks very much the same for all scenarios, with the exception of the case with $\nu = 5$ and $n = 1000$, in which case there is strong information in the data to dominate the very informative exponential priors.

Another very insightful way to look at Figure 7 is column-wise. If we look at the first column, from top to bottom, we clearly see the difference in the posterior inference obtained with the divergence priors when we go from Gaussian data, to more heavy-tailed data. That is, the medians and the interval end points get more and more concentrated around low values of ν. The same pattern happens even more clearly when the number of data points is increased to $n = 1000$, which can be seen in the third column of Figure 7. Now, if we look at the second column, from top to bottom, we basically see no difference in posterior inference when the exponential priors are used for ν, even when we compare the scenario with Gaussian data with that of strong heavy-tailed data, that is $\nu = 5$. This means that in this case the results will be much more influenced by the choice of λ for the exponential distribution than by the kind of data you have, which is definitely not a good sign. A similar pattern happens for the first two rows in the fourth column, which represents the scenario with $n = 1000$. The only exception happens on the last row of the fourth column. When $\nu = 5$, the data carries enough information about the tails of the distribution to convince the models with exponential distribution for ν that the tails of the data are indeed very different than the Gaussian tails. But looking at the last row of Figure 7, our priors have detected this even with $n = 100$, while the exponential priors required $n = 1000$ to get convinced about the huge difference between a Gaussian dataset and a Student's t dataset with $\nu = 5$. This kind of dynamic behavior observed when the number of data points increases is better illustrated by looking at Figure 8.

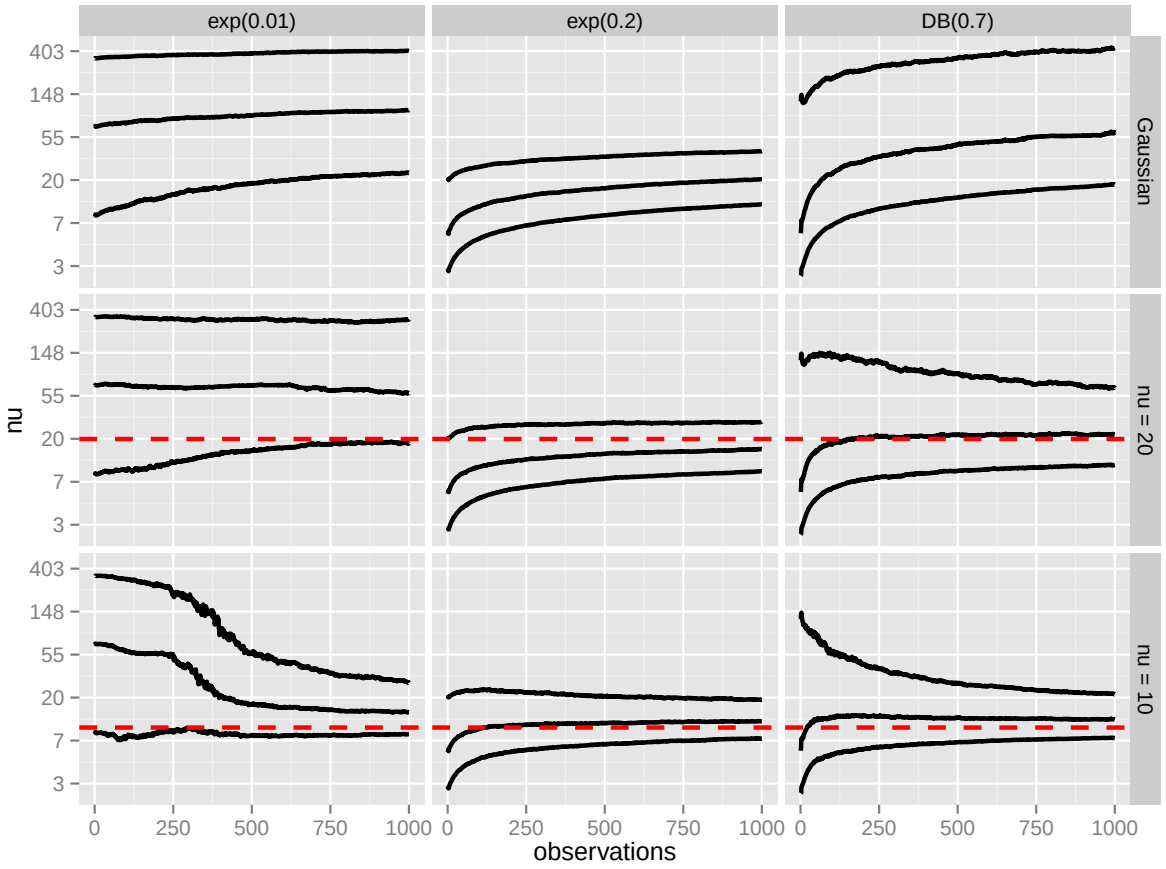


Figure 8: Median of the $(0.025, 0.5, 0.975)$-quantiles of the posterior distribution of ν obtained from sequentially fitting the robust linear regression model with different prior distributions for ν to 1000 datasets where the posteriors were computed for each $n = 1, 2, ..., 1000$, given all the previous observations $y_{1:n} = \{y_1, ..., y_n\}$. This sequential analysis were computed for degrees of freedom $\nu = \infty, 20, 10$ and three different priors for ν, the exponential prior with rate parameters $\lambda = 0.01$ (left), $\lambda = 0.2$ (middle) and with our divergence-based prior with df $= 0.7$ (right).

Figure 8 illustrates a different representation of the same problem, where the median of the $(0.025, 0.5, 0.975)$-quantiles of the posterior distribution of ν obtained from fitting the models to 1000 datasets were computed for each $n = 1, 2, ..., 1000$, given all the previous observations $y_{1:n} = \{y_1, ..., y_n\}$. It presents this sequential analysis for the robust linear regression models with three different priors for ν and data simulated from the model with $\nu = \infty, 20, 10$. The first column shows the results obtained with the exponential prior for ν with $\lambda = 0.01$ which gives satisfactory results for Gaussian data but not so for lower degrees of freedom, unless a lot

of data is provided, even for a simple model like the one considered here. The second column illustrates the results from the model with exponential prior for ν with $\lambda = 0.2$, which shows (a misleading) good performance for low degrees of freedom, as we can see for the case with $\nu = 10$ where the model gives credible intervals tightly concentrated around the true ν even for very few observation, illustrating that the prior itself is concentrated around this low value of ν. However, this second model is also reasonably sure that ν is low even in the presence of Gaussian data, being necessary much more data points to convince it otherwise. In contrast, the divergence prior with df $= 0.7$ behave in accordance with the framework described in Section 3.2. We see a better representation of our uncertainty in the presence of few observations with wider credible intervals and fast convergence towards the true value of ν as more data arrives. This happens for every scenario, from high to low degrees of freedom, so that we see no reason why not to use this class of priors, unless there is a strong belief about the value of ν, in which case might be more appropriate to use a prior that encode this information.

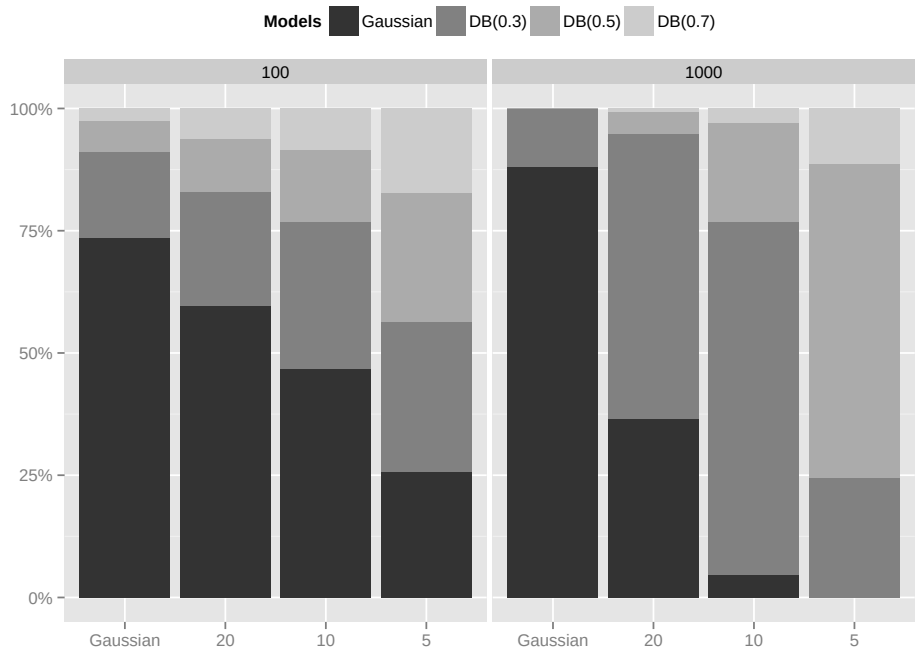


Figure 9: Frequency that each of four models, namely linear regression model and robust linear regression models with divergence-based priors having df $= 0.3, 0.5, 0.7$, have been chosen according to the Bayes factor computed for 1000 datasets for 8 different scenarios with sample size $n = 100, 1000$ and degrees of freedom $\nu = \infty, 20, 10, 5$.

By construction, our framework provides an intuitive way to try different priors for the flexibility parameters. By increasing the df parameter, the user knows what is happening, namely that he is increasing the flexibility of the model without forcing it to be non-Gaussian. Besides, among many options, Bayes factor can be used to decide what is the degree of flexibility required for each problem. Figure 9 provides the frequency that each of four models have been chosen according to the Bayes factor computed for 1000 datasets for 8 different scenarios. The models considered in Figure 9 are a simple linear regression model and the robust linear regression models with divergence-based priors having df $= 0.3, 0.5, 0.7$. There were eight scenarios considered with sample sizes $n = 100, 1000$ and degrees of freedom $\nu = \infty, 20, 10, 5$. The results are as expected. Even though we have shown that our flexible models give good results in case the data is Gaussian, it is still preferable (based on Bayes factor) to use a Gaussian model if the data is Gaussian. A model based on the Student's t becomes preferable as the tail of the data becomes heavier, and this effect gets stronger as we increase the sample size. Lastly, according to Figure 9 we note that more flexible models, as for example with df > 0.5, are indicated only for very heavy-tailed data. Note that Hjort (1994) have shown that if $\nu \geq 1.458\sqrt{n}$, then maximum likelihood estimation using an incorrect Gaussian model is still more precise than using the correct model based on Student's t assumptions.

More than highlight the disadvantages of the exponential prior for the degrees of freedom ν of the Student's t distribution, this example is important to illustrate a very desirable property of the priors designed with our framework. The divergence priors perform well if the data are close to the basic model, i.e. the Gaussian distribution in this context, while being able to capture deviations from this basic model if the data says so. In addition, even though the degrees of flexibility parameter, df, influences the posterior inference, Figure 7 shows that posterior inference are quite robust to the specification of df. This means that the most flexible prior considered in this experiment, with df $= 0.7$ still fares quite well in the presence of Gaussian data, while the less flexible one, with df $= 0.3$ also gives sensible results in the presence of heavy-tailed data. Besides, we have shown that our priors makes the task of increasing and decreasing the flexibility of a given model more intuitive.

5.2 Stochastic volatility example

In order to show that the behavior observed in Example 5.1 remains valid under more complex models, we compare the performance of the divergence-based priors and the exponential priors for the degrees of freedom of the stochastic volatility model. The observational equation is given

by

$$y_t = \sqrt{\exp(f_t)}\varepsilon_t, \quad \varepsilon_t \sim T_\nu(0,1),$$

where $T_\nu(0,1)$ is a standardized Student's t distribution with mean 0, variance 1 and degrees of freedom equal to ν, as described in Section 3.1. The time-dependent variance of y_t is then given by $\exp(f_t)$. The log of the variance f_t will follow an auto-regressive process of order 1 (AR1) defined by

$$\begin{aligned} f_t &= \rho f_{t-1} + w_t, \quad w_t \sim N(0, \tau^{-1}), \quad |\rho| < 1 \\ f_1 &\sim N(0, (\tau(1-\rho^2))^{-1}), \end{aligned}$$

where a gamma prior will be assigned to the marginal precision of the AR1 process, $\kappa = \tau(1-\rho^2)$ and a Gaussian prior will be assigned to $\phi = \log\left(\frac{1+\rho}{1-\rho}\right)$.

We will analyse two datasets, the first is formed by weekly observations of the S&P index (SP500) from 01/06/1984 to 30/06/2013 and the second by daily observations of the SP500 from 01/06/2007 to 30/06/2013. Those dates were chosen so that each dataset has approximately 1500 observations. But before going to a real dataset analysis, we performed a similar simulation study that originated Figure 8 in Example 5.1 for the stochastic volatility model just to make sure we could expect similar behavior on the inference about ν as the one found in Example 5.1. We have simulated 200 datasets each with 2000 data points under three different scenarios $\nu = \infty, 20, 10$. We have used $\rho = 0.95$ and $\tau = 16$ to mimic similar values found on the analysis of the SP500 datasets. Figure 10 shows similar behavior to that obtained in Figure 8. The divergence-based prior did well across different scenarios, while the exponential priors demonstrated again a huge sensitivity with respect to its hyperparameters, performing well only on the scenarios they were designed to work and poorly elsewhere.

We analyse both weekly and daily SP500 data because there is evidence in the literature (Jacquier et al., 1994) that daily data tends to have heavier tails when compared to weekly data and we think it would be interesting to see if our models could capture this behavior and outperform the models based on the exponential priors. We have used continuous return $r_t = \log(P_t/P_{t-1})$ as a response variable, where P_t is the SP500 index at time t. Figure 11 shows both the weekly and the daily standardized datasets.

Figure 12 shows the $(0.025, 0.5, 0.975)$-quantiles from the posterior distribution of ν for the weekly (first row) and daily (second row) SP500 data using the first 500 data points (left column), first 1000 data points (middle column) and finally all 1500 data points (right column). For each scenario we have 7 credible intervals, obtained by using different prior distributions for ν. The first three solid intervals use our divergence-based priors with $\text{df} = 0.3, 0.5, 0.7$ respectively. The

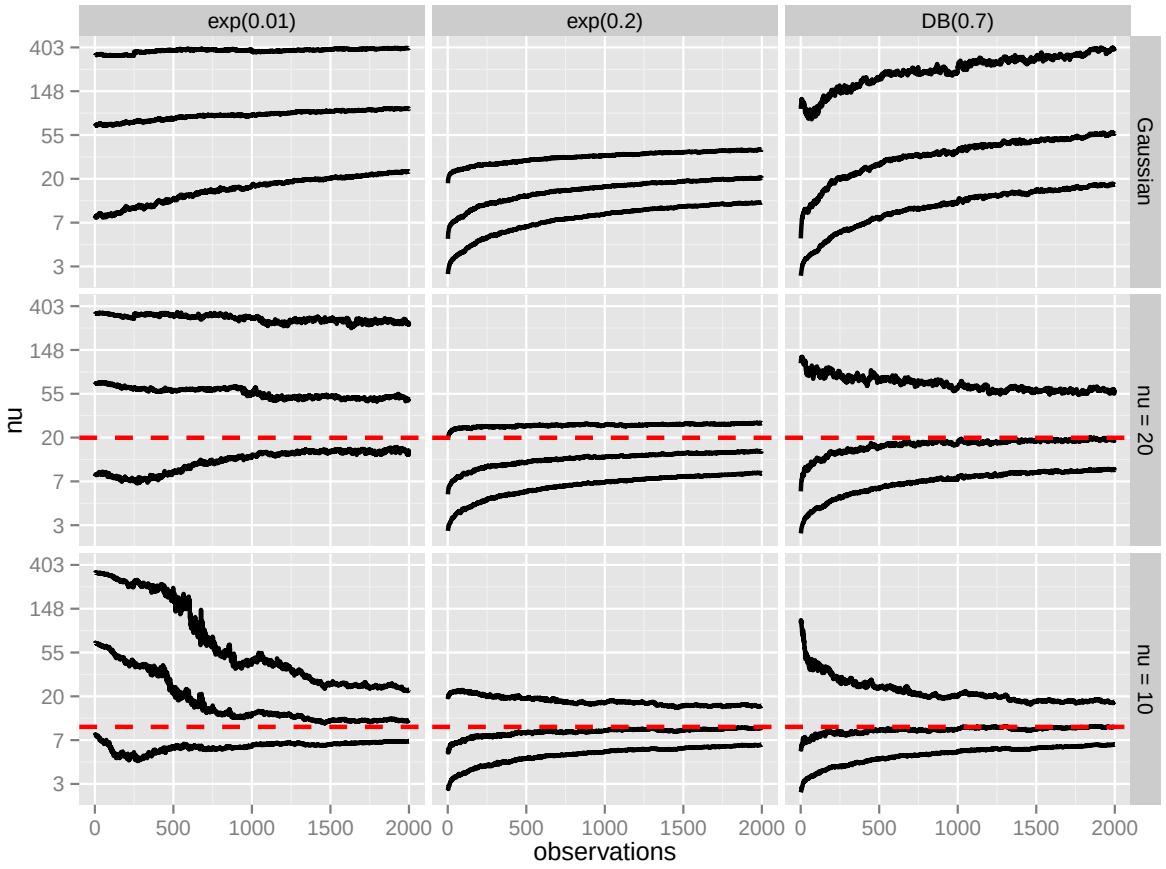


Figure 10: Median of the $(0.025, 0.5, 0.975)$-quantiles of the posterior distribution of ν obtained from sequentially fitting the stochastic volatility model with different prior distributions for ν to 200 datasets where the posteriors were computed for each $n = 1, 2, ..., 2000$, given all the previous observations $y_{1:n} = \{y_1, ..., y_n\}$. This sequential analysis were computed for degrees of freedom $\nu = \infty, 20, 10$ and three different priors for ν, the exponential prior with rate parameters $\lambda = 0.01$ (left), $\lambda = 0.2$ (middle) and with our divergence-based prior with df $= 0.7$ (right).

last four dashed intervals use exponential priors with $\lambda = 0.01, 0.05, 0.1, 0.2$, respectively. Again, we can see that our divergence-based priors have demonstrated low sensitivity with respect to changes in the hyperparameter df and successfully detected a higher degrees of freedom scenario for the weekly data when compared to the daily data, which was expected given the nature of the data. The models based on the exponential prior for ν are again very sensitive to the choice of the hyperparameter λ of the exponential distribution and the conclusions regarding the inference about ν were more influenced by the choice of λ than by which data, weekly or

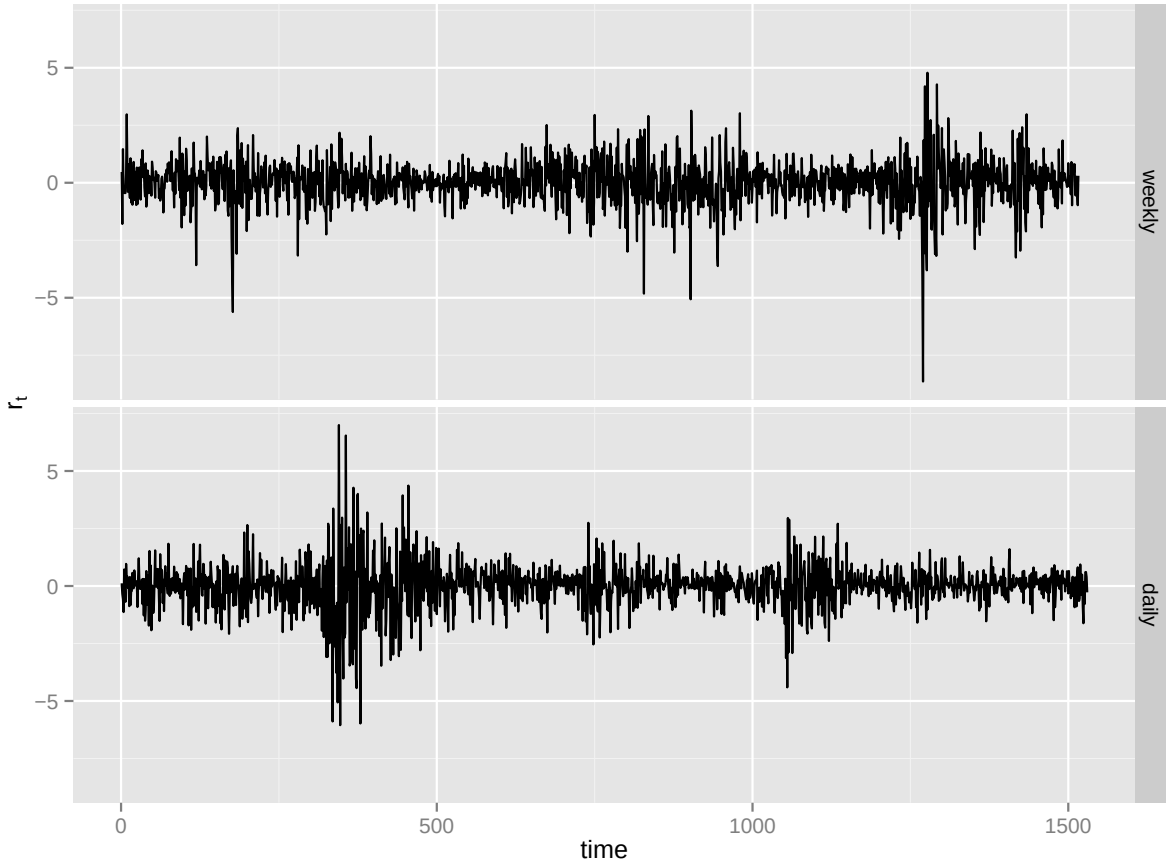


Figure 11: Weekly and daily standardized continuous returns of SP500 index. The weekly data were observed from 01/06/1984 to 30/06/2013 and the daily data were observed from 01/06/2007 to 30/06/2013. The dates were chosen so that each dataset has approximately 1500 observations.

daily, the models were exposed.

6 Conclusion

In a typical data analysis context, it is not uncommon to start with a basic model, usually well established in the literature, and then expand it to form a more flexible model to account for possible deviations from the basic model when supported by the observed data at hand. One way to expand the initial model is to use a parametric family that contains the basic model as a particular case, and have its flexibility controlled by a given parameter, which we have denoted by flexibility parameter. We have developed a formal framework to construct prior distributions

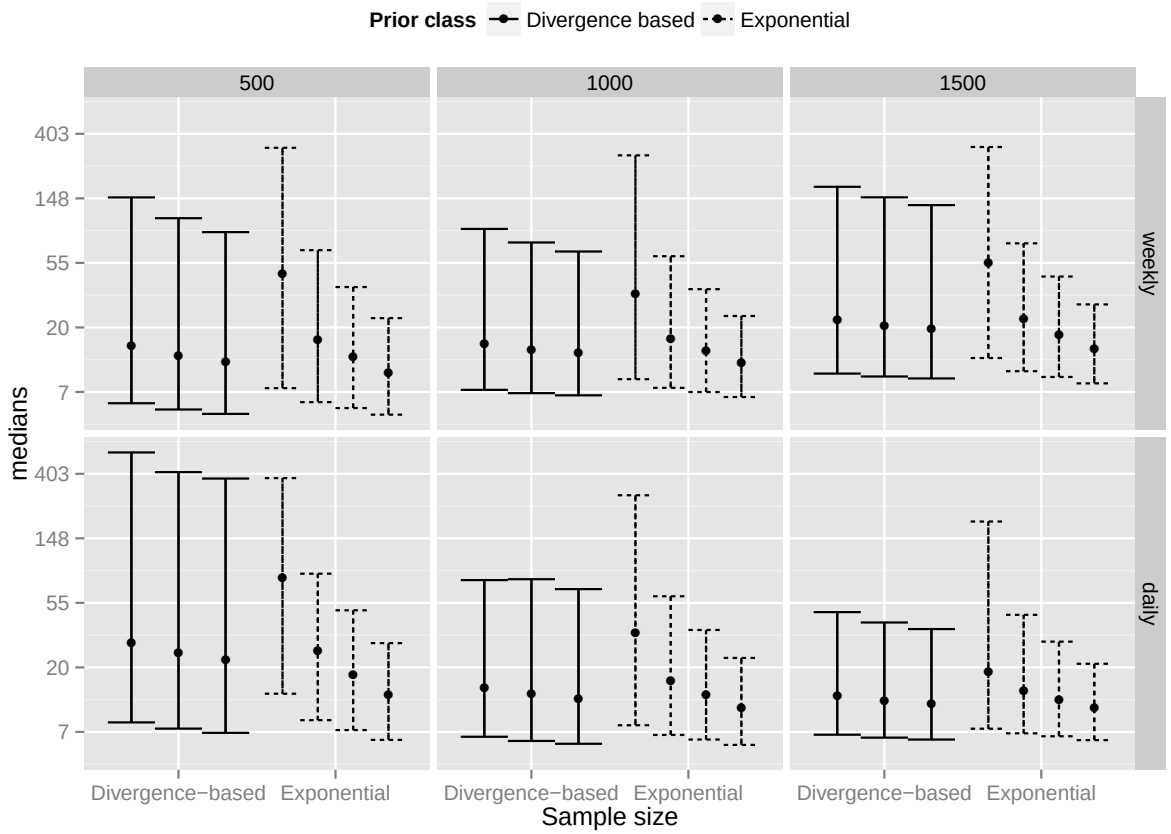


Figure 12: $(0.025, 0.5, 0.975)$-quantiles from the posterior distribution of ν for the weekly (first row) and daily (second row) SP500 data using the first 500 data points (left column), first 1000 data points (middle column) and finally all 1500 data points (right column) using our divergence-based priors with $\text{df} = 0.3, 0.5, 0.7$ (solid intervals) and exponential priors with $\lambda = 0.01, 0.05, 0.1, 0.2$ (dashed intervals).

for the flexibility parameter that place the basic model in a central position within the flexible model, and allow the user to intuitively control the amount of flexibility around the basic model. We assign the prior distribution on the scale of the divergence between the basic and flexible model and then map it back to the original scale of the flexibility parameter. This allow us to properly encode our assumptions on the prior distribution of the flexibility parameter.

We have illustrated the application of our framework in the case where we want to relax the Gaussian assumption to allow heavier-than-normal tails by using a Student's t distribution, in which case the flexibility parameter is the degrees of freedom of the Student's t distribution. In

this context, we have shown that our priors are robust with respect to its hyperparameters and give sensible results across many different scenarios. We have also shown disadvantages of using priors that do not place the basic model in a central position within the flexible model, which implies that the models built upon those priors are much more sensitive to misspecification of the hyperparameters of the prior for the flexibility parameter. The shortcoming of those priors can be better understood if we map them to the divergence scale between the basic and the flexible model. Neither the divergence measure nor the prior assigned to the divergence scale are unique, but we have suggested the use of the calibrated Kullback-Leibler divergence and the exponential distribution as the divergence measure and prior distribution assigned to the divergence scale, respectively.

Although we have used the Student's t case to motivate and illustrate our framework, it is important to note that the framework presented here is applicable to any model that can be seen as an extension of a basic model. For example, a spline model can be seen as an extension of a straight line model and the smoothness parameter of the spline model would control the deviation from the straight line and could be seen as a flexibility parameter in our framework. Another example is when a spatial model can be seen as an extension of an independent components random effects model, and the parameter controlling the strength of the spatial dependence would control the deviations from the independent component model and could be seen as the flexibility parameter in our framework. A detailed description of the application of our framework to those two cases mentioned above is part of our current research agenda.

References

Bayarri, M. and García-Donato, G. (2008). Generalization of jeffreys divergence-based priors for bayesian hypothesis testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):981–1003.

Box, G. and Tiao, G. (1964). A bayesian approach to the importance of assumptions applied to the comparison of variances. *Biometrika*, 51(1/2):153–167.

Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain monte carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316.

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Fonseca, T., Ferreira, M., and Migon, H. (2008). Objective bayesian analysis for the student-t regression model. *Biometrika*, 95(2):325.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383.

Geweke, J. (2006). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40.

Hjort, N. L. (1994). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89(426):665–675.

Jacquier, E., Polson, N., and Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic statistics*, 20(1):69–87.

Jacquier, E., Polson, N., and Rossi, P. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics*, 122(1):185–212.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Jeffreys, S. (1961). *Theory of probability*. Oxford University Press, USA.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lange, K., Little, R., and Taylor, J. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, pages 881–896.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067.

Masreliez, C. and Martin, R. (1977). Robust bayesian estimation for the linear model and robustifying the kalman filter. *Automatic Control, IEEE Transactions on*, 22(3):361–371.

Pinheiro, J., Liu, C., and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10(2):249–276.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 431–439.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.